LogForum

> Scientific Journal of Logistics <

2022, 18 (3), 357-364

http://doi.org/10.17270/J.LOG.2022.738

http://www.logforum.net p-ISSN 1895-2038 e-ISSN 1734-459X

ORIGINAL PAPER

IDENTIFYING THE COGNITIVE GAP IN THE CAUSES OF PRODUCT NAME AMBIGUITY IN E-COMMERCE

Maciej Niemir[1], Beata Mrugalska[2]

1) Łukasiewicz - Poznań Institute of Technology, Faculty of Engineering Management, Poznan University of Technology, Poznan, **Poland**
2) Faculty of Engineering Management, Poznan University of Technology, Poznan, **Poland**

ABSTRACT. Background: Global product identification standards and methods of product data exchange are known and widespread in the traditional market. However, it turns out that the e-commerce market needs data that have not already received much attention, for which no standards have been established in relation to their content. Furthermore, their current quality is often perceived below expectations. This paper discusses the issues of product name and highlights its problems in the context of e-commerce. Attention is also drawn to the source of liability for erroneous data.
Methods: The research methodology is based on the analysis of data of products available on the Internet through product catalog services, online stores, and e-marketplaces, mainly in Poland, but addresses a global problem. Three research scenarios were chosen, comparing product names aggregated by GTIN, starting with e-commerce sites and ending with product catalogs working with manufacturers. In addition, a scenario of name-photo compatibility was included.
Results: The results show that the product name, which in the real world is an integral part of the product as it appears on the label provided by the manufacturer, in the virtual world is an attribute consciously or not modified by the service provider. It turns out that discrepancies appear already at the source - at the manufacturer's level - publishing different names for the same product when working with data catalogs or publishing on product pages contributing to the so-called snowball effect.
Conclusions: On the Internet, products do not have a fixed name that fully describes the product, which causes problems in uniquely identifying the same products in different data sources. This in turn reduces the quality of data aggregation, search, and reliability. This state of affairs is not solely the responsibility of e-commerce marketplace vendors, but of the manufacturers themselves, who do not take care to publicize the unambiguous and permanent name of their products in digital form. Moreover, there are no unambiguous global guidelines for the construction of a full product name. The lack of such a template encourages individual interpretations of how to describe a product.

Keywords: basic product data, product catalog, e-data catalog, master data, data quality, dirty data

INTRODUCTION

In recent years, the popularity of e-commerce has increased significantly. Moreover, the COVID-19 pandemic has further accelerated it [Lone et al., 2021]. Unfortunately, it is noted that many companies still decide to obtain product data by copying them from other e-stores, downloading them from social databases available on the Internet with an untested source and in a different format, or writing data from product labels according to their own guidelines. Furthermore, companies, competing for customers on the Internet, also decide to introduce their modifications of product attributes, including photos and product names, trying to attract the attention of a potential buyer. It shows how e-commerce produces poor quality data, which is the so-called "dirty data" [Guoling & Qinyun, 2008]. "Dirty data" is a phrase that denotes out-of-date, inconsistent, or incomplete data [Zhou et al., 2011] that cannot be properly searched, aggregated, and analyzed. Operations on such data generate costs within the organization [Haug et al., 2011], and also reduce the credibility of the company in terms of the products sold [Qalati et al., 2021]. In addition, inconsistencies in the data are one of the causes of product returns, which, in turn, due to the need

Citation: *Maciej Niemir, Beata Mrugalska, 2022. Identifying the cognitive gap in the causes of product name ambiguity in e-commerce. LogForum 18 (3), 357-364,* http://doi.org/10.17270/J.LOG.2021.738
Received: 20.02.2022, Accepted: 20.09.2022, on-line: 30.09.2022

to organize additional transportation, negatively affects the environment [Kawa & Pierański, 2021].

This paper aims to present the quality issues related to product names in e-commerce. It shows the scale of the problem, examples of occurrence, possible reasons why such errors occur, and who is responsible for them. The intention is to draw attention to the problem of the content of text data in product catalogs and to the consequences of the lack of guidelines in this regard and, as a result, to initiate a discussion on the need for standardization in this area.

The paper is prepared as follows: Chapter 2 provides a review of the literature on data quality. Chapter 3 shows the meaning of the product name, which is one of the most important attributes of the product catalog. Chapters 4, 5, 6, and 7 present the research results and discussion. Finally, Section 8 provides a summary of the conclusions.

IMPORTANCE OF PRODUCT DATA QUALITY

In general terms, data quality is described as the ability of data to satisfy stated and implied needs when data are used under specified conditions [ISO 25012]. In the literature, data quality is often defined as "fitness for use" [Wang & Strong, 1996], [Batini et al., 2009] and it is understood as the ability of a data collection to meet users' requirements. As early as 1968, it was found out that the desired attributes defining quality are relevance, timeliness, and accuracy [Feltham, 1968]. As noted by Reeves & Bednar [1994], there are five definitions related to the concept of quality: value, compliance with specifications, compliance with requirements, fitness for use, and meeting and/or exceeding customer expectations. The concept of quality is usually considered as a multidimensional concept, for example, Wand & Wang [1996] proposed the decomposition of data quality into four internal dimensions: complete, unambiguous, meaningful and correct, while Cichy & Rass [2019] identified twenty of them. When we consider high quality, the data/information must be relevant, accurate, factual, complete, reliable, structured, precise, readable, and reasonable [Zmud, 1978].

In the case of e-commerce, the identification of the needs of product data is an issue that should be considered broadly and in a proper perspective. Product data, used locally on the e-shop website, may be used for other purposes in the long run. They can be used to synchronize and integrate, such as with price comparison sites, e-marketplace platforms, or even to increase the visibility of an e-store by search engines through the use of structured data, but they can also be used to exchange information on a logistical level. The lack of standardization in such terms will completely block the possibility of using modern technologies and minimizing the so-called "Human Factor" [Żuchowski, 2022]. It is worth emphasizing it even if the company does not see such a need at present and the scope of data is considered sufficient to conduct its business. If, after some time, it turns out that the data are not complete - that is, it does not meet the requirements of business partners in the minimum dimension for integration [Niemir & Mrugalska, 2021] (for, e.g., the set does not have GTIN numbers, there are no data on the product brand or the data are disordered – wrong format of product name or wrong format of main photo is used), it will certainly be a significant barrier or even prevent cooperation due to the costs of changes in the product database. Marsh [2005] confirmed it in his research, where due to "dirty data" 88% of data integration projects completely failed or exceeded the assumed budget, while 33% of organizations delayed or withdrew from IT implementations.

MEANING OF PRODUCT NAME

In traditional trade and communication, a key product attribute is GTIN (Global Trade Item Number) which uniquely identifies the product, while in e-commerce, product identification begins with its name [Niemir & Mrugalska, 2021]. For example, in a stationary store, the customer identifies the product by packaging, and the purchasing process begins with scanning of the bar code, i.e. GTIN identification, the IT system uses it to determine the price, stock levels, etc. The name of the product appears only on the receipt and in many cases, it does not identify the product and is only illustrative. In the case of online purchases, the product is identified and searched for by the name. The GTIN does

not matter at this stage, and the purchase takes place at the level of the internal e-shop ID associated with the name. It shows why the product name is so important. Unfortunately, it turns out that just as the GTIN number is a permanent element, given by the manufacturer, the name of the product is interpreted by everyone in their own way.

MATERIALS AND METHODS

The research methodology was based on the analysis of data from product databases from various sources and was divided into four different scenarios. The first scenario referred to the study of the problem on a large scale. Three different product catalogs were compared in terms of the similarity of product names for products with the same GTINs in each database. We have selected product catalogs that were created in cooperation with manufacturers and were used to provide data-sharing services. This choice aimed to prove that this quality problem is common and occurs at the very source. The other scenarios were already case studies. In the second scenario, we focused on information from e-stores, knowing that in many cases their owners obtain product data on their own. It allowed us to verify to what extent their interpretation of the product name corresponds to reality. After downloading product data from several e-stores (web scraping), they were compared with each other, and one of the most frequently repeated, popular products were selected for further analysis. The third scenario was developed on the basis of data from the e-marketplace, i.e., a platform associating producers, distributors, and sellers, where data of offered products are entered through a common tool. E-marketplaces are places where problems with product names and their aggregation are particularly visible. Their customers are often small, inexperienced sellers. In such a case, it is easy to imagine the consequences of misunderstanding the need to enter the name of the product instead of the name of the offer. The largest Polish platform of this type was selected for the purpose of the research, and the product was selected on the basis of the available summary of data aggregated according to GTIN. The last scenario concerned a deeper problem as it was related to the interpretation of the text. It showed how the name could contribute to misconceptions about the product. The product for the analysis was selected experimentally on the basis of the authors' observations, using the Google Search Engine.

PRODUCT NAME IN SELECTED E-STORES

The study of the quality and consistency of product name data began with an analysis of the offers of dozens of online stores offering FMCG (Fast Moving Consumer Goods) products to the Polish market. The acquired data was aggregated using GTIN (Global Trade Item Number, a unique product number encoded in the form of a barcode). Of repeated product listings with the same GTIN numbers in different stores, 100 of them were selected, whose numbers were registered in different countries. Germany, UK, Greece, Portugal, Poland, Hungary, Sweden, Switzerland, Italy, Netherlands, and Austria. Products included beverages, candy, cleaning products, and cosmetics. Then, it was examined whether the product names were identical across the various offerings. The result was clear: for each of the selected products, their names were different in different stores.

To analyze the problem in detail, a popular Polish food product was selected from among the products that had an unusually large number of name variations. Its manufacturer maintains the official website, of the product on its website but without information about the GTIN number and the full name that describe this product. The header of the page (the leading H2 header of the page) reads "Lemon Flavor", and there is also a photo of the product. The product is registered in the official GS1 Polska register as "Lemon cake flavor 9mlx20pcs EXP". For research purposes, the manufacturer's brand name was anonymized with the value "[BRAND]", leaving the original letter size. The names of this product were originally in Polish, but they were translated without losing the context. The results obtained are presented in Table 1.

It needs to be underlined that under this GTIN there is a single product, not 20 items (the collective packaging has a different GTIN number) and the net content of the product is 9 ml, not 10 ml, 9 Ml (mega liter?) or 10 g. During the additional search of websites by product

name, an identical product with a wrongly marked GTIN number was also found. From the point of view of Internet search engines, especially intelligent ones, the terms, such as "glass" and especially "bot", probably an abbreviation for "bottle", are very misleading information. It adversely affects the accuracy of the results. For NLP / AI algorithms, in addition to the above, the inflection of the word and its use in context also play a role. One thing is "lemon flavor", and another thing is "lemon fruit".

Table 1. Different names for an example product.

[BRAND] LEMON FLAVOR 10ML	FLAVOR [BRAND] CITR 9ML
[BRAND] Lemon Flavor 9 Ml	flavor [Brand] lemon
[BRAND] Lemon cake flavor	Flavor [Brand] Lem 9ml [Manufacturer]
[Brand] Lemon cake flavor 10 ml	LEMON FLAVOR 10ML [BRAND]
[Brand] lemon cake flavor 10ml bot	LEMON FLAVOR
[Brand] Lemon cake flavor 9 ml	LEMON FRUIT FLAVOR [BRAND]
[BRAND] Lemon Cake Flavor 9 Ml	LEMON FLAVOR 10ML * [BRAND] * EU
[BRAND] Lemon cake flavor glass	LEMON FLAVOR 10ML [BRAND]
[Brand] flavor for cakes, creams, and punch, lemon 10 ml	LEMON FLAVOR 9ML [BRAND]
[Brand] Lemon Cake Flavor 10g.	CAKE FLAVOR [BRAND] 9ML LEMON FRUIT IO
[BRAND] Lemon Aroma for cakes, creams and punch 10ml	LEMON CAKES FLAVOR [BRAND] 9ml
[BRAND]: LEMON FLAVOR	LEMON CAKES FLAVOR 9ML [BRAND]
[BRAND] -LEMON CAKE FLAVOR 10G A18.	

Source: Own work.

From the analyzed example, it is possible to observe quite common differences in product names, probably due to the fact that e-commerce database operators enter data from product labels without any guidelines or under their own internal regulations. This situation causes that the names of the same product are different each time: the text starts with either brand or common name, purpose, etc. ending with net content, type of packaging, or internal store markings. The research additionally shows that in some stores information about the product is incomplete or incorrect, which is an even more serious problem. This detailed case also made it possible to observe that the quality problem arises not only at the retailers' level, but also already at the manufacturer's level. The product data on its official product website is incomplete, while the registered data in GS1 Poland assigned to this GTIN is incorrect.

Due to doubts that the problem of quality and differences in names does not lie in the specifics of the local market and language, name analysis was also carried out for several products available on the German and British markets while maintaining the same research methodology. The results were identical.

Comparing the situation to a traditional purchase - this problem will not occur - we buy a product that we can see. The label of most of products is usually designed to make the brand and the main characteristics of the product easy to read. The customer should not be misled. Also, the label of the product is the same in each store because it comes from the manufacturer. As long as the GTIN number is on the product label and in the database of the store's IT system, the number is assigned appropriately - there are no problems with the purchase.

PRODUCT NAME / OFFER IN THE E-MARKETPLACE

In further research, an observation of a product with the same GTIN entered for multiple offer names and categories was made on a Polish shopping platform. It is an example where a buyer, trying to search for different products, receives multiple offers with the same product. The platforms currently try to solve such problems by creating their own product catalogs (Amazon, eBay, Allegro), aggregating such data

by GTIN number, and then displaying the product using the agreed common product name. However, in spite of the fact that it is one of the most interesting ways to solve such a problem, it will be effective at a local level – in an e-marketplace. Additionally, it will be only the solution if the platform has reference data and the same GTIN is present in all aggregated data sources. In practice, unfortunately, the solution is not 100% effective.

For a selected exemplary product, a total of 771 pairs were calculated for the set: offer name (product) + product category. The product appeared in 10 different categories and 283 different names, and the data described the toy as a microphone with a battery-powered speaker and the possibility of connecting an additional audio source via Bluetooth and having lighting effects. Depending on the trader and a specific offer, the product was classified as: “light gadget”, “electronic gadget”, “portable speaker”, “karaoke equipment”, “wireless microphone”, “GSM accessory” and finally as “a toy”. Instead of the name of the product, the offer contained various information, from the common name to the description of the customer's needs. For example: “for a girl”, “to the phone [provided phone model name]”, “karaoke glowing LED”, “as a gift for CHILDREN”, “light and loud LED lights”, “to party”, “Father's Day gift idea”.

The example shows the situation where the seller uses the attribute of the offer to attract attention at the cost of reliable information about the product, just to increase the sales. The selected product did not have a positioned brand or even a recognizable product name. Its producer was probably just entering the market. The described situation could temporarily give the seller profits but globally aggravates the problem of data quality. The e-marketplace also loses due to such activities, several dozen or more pages with the same product under different names of offers certainly does not encourage the buyer to continue the search. This action does not also position the brand, manufacturer, or product. However, similar scenarios are impossible in traditional commerce.

MISCONNECTION OF NAME AND PHOTO

Complex and multi-criteria problems can appear when we comply a product name with its photo. For example, such situations can result from color interpretation. In order to check it, "woman mint pants" were entered in the Google search engine. The data was obtained from e-stores where a photo was attached to a product and a name including the above phrase was inserted. In the study, all stores were analyzed in detail to confirm the word "mint" in relation to the name to eliminate situations when the search engine retrieved data, e.g. from the description of the product page. The search results are presented in Figure 1 in the same order as it was displayed on the page.

In order to start the analysis of this case, two questions can arise: "what color of pants can you see in the photos?" or "what color of pants will you get as a customer"? Celadon, turquoise, or mint? We can differentiate many reasons for this situation such as incorrectly entered product name by the operator of the online store database or badly loaded photo that does not represent the actual product or even photo with poor quality, e.g. distorted colors, wrong color definition. Of course, one can ask the question who introduced such a name? An e-shop employee or a manufacturer of the cloth?

Another problem is the photo itself - the shot, the background, the number of products in a single photo as well as the presentation of the model or the cloth itself. These are problems of data standardization, which is lacking not in terms of the field name that defines the photo, but in terms of the content [Niemir & Mrugalska, 2022]. From the photo, you cannot read what the subject of the purchase is - pants, a set of pants + blouse, handbag, or shoes. As was mentioned before it is also not known what will be the color of the product when it is finally delivered.

Today’s e-consumers are accustomed to situations where the filtered results do not correspond to their questions. In comparison to a purchase in a stationary store, such situation will never occur as we buy the real product which we can consciously see, touch and sometimes even feel and smell.

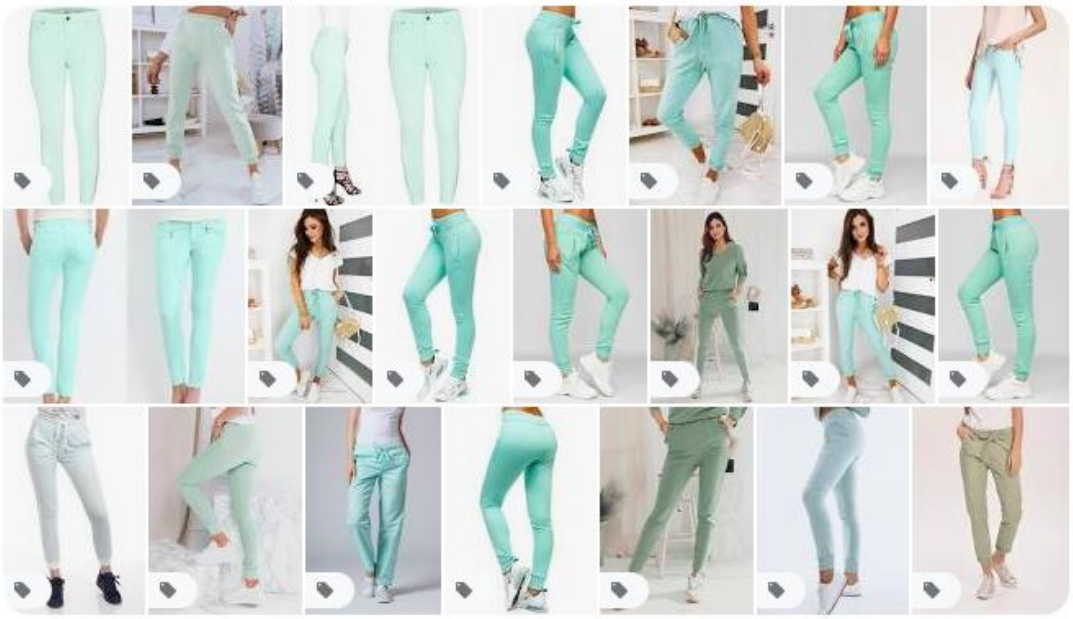

Fig. 1. An example of photos of products for the searched phrase "Woman mint pants ".
Source: Google Image Search.

SIMILARITY OF NAME IN PRODUCT CATALOGS

Due to previous research results, which proved that in some cases ambiguities arise at the very source of the data, i.e. the manufacturer, it was decided to perform an additional study to confirm this thesis. Three data catalogs that provide data services for the e-commerce market, whose information is verified in cooperation with manufacturers, were selected. The research was conducted for products appearing at once in all three catalogs, i.e. having the same GTIN. An additional selection criterion was the availability of a given product on the Polish market, regardless of its origin and the country of the manufacturer. Full aggregation was obtained for 9266 products surveyed. A test was then carried out to see if the name of a given product was the same (exact match of text strings without case comparison) and, if not - whether it was similar. The similarity of the names was verified with an algorithm that ignored differences in the order of words/letters and numbers in the name, while the number of letters and numbers in both cases had to match. The results of this study are shown in Figure 2.

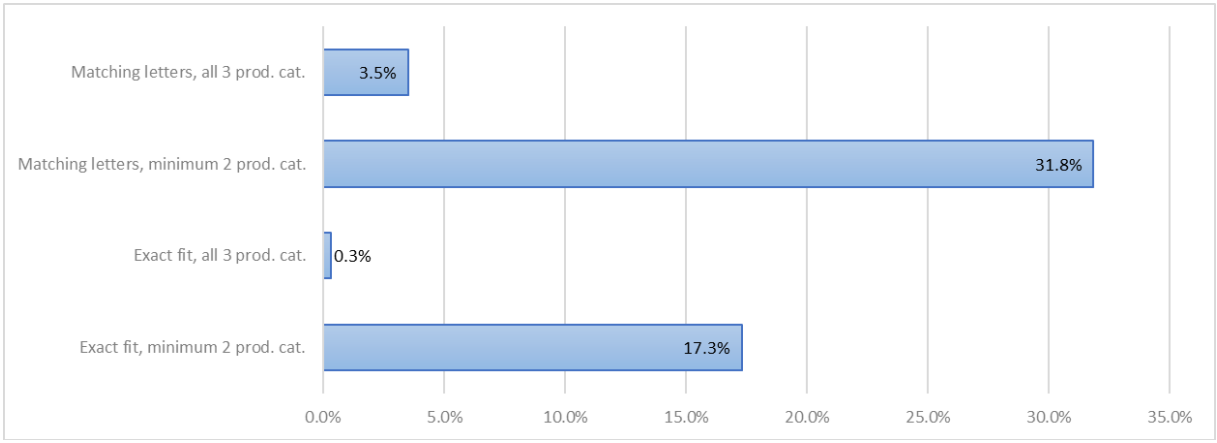

Fig. 2. Conformity of product name.
Source: own work.

The complete concordance of the product names in all 3 catalogs practically did not occur, the concordance was recorded for only 30 of them. Omitting one catalog, the concordance reached only 17.3% of the total set. The study of similarities, comparing the occurrence of the same letters and numbers with the same number of them in both names, gave better results, although the study itself should be treated as an overview, due to the flawed nature of the algorithm. Despite this, the similarity of the names still remained low.

CONCLUSIVE REMARKS

The lack of clarity in the names of products sold via the Internet is a significant problem, especially when we take into account data aggregation, data verification with the real product, and data search by phrase. In the virtual world, the basic attributes of products, especially the name and photo, which directly reflect reality, should not be freely interpreted. As in the GTIN, the basic product data should derive from the manufacturer and be closely related to the product - be "stuck to the product" just as a label. It is not that only retailers should care about the quality of the description of the products they offer, but it is the manufacturers, who should describe the attributes in a standardized way and publish the data associated with the product in such a way that anyone can easily read it and verify its accuracy in their own database, which, however, is not currently happening. The results of this research clearly show that the data can be incorrect or incomplete at the source, further increasing the number of errors and even compounding them throughout the supply chain. Particularly in e-commerce, when we do not see the product being sold, but only its data, the product name should contain fully reliable information that strictly describes the specific product and should unambiguously identify it, just like a product label in the real world. Therefore, the product name should contain at least such data as the brand name, common name, product variant, and net content. There is no chance of correct verification and unambiguous traceability of the product without such information arranged in the right order. Similarly, the photo of the product - the one that is the main and most important - should reflect the actual purchase. So, if the purchase relates to one product, the photo should only present that specific product. If such a product is, for example, "pants", then the main photo should not be a drawing or a sketch, be presented on a model, be presented in a specific scenery, etc. The standardization of this area is not easy because the problem affects all products in all industries, and each restriction in both the text of the product name and the presentation of the product as the image may be difficult in specific situations. Nevertheless, it is worth taking steps to standardize the data and taking the responsibility, as a manufacturer, for their correct definition and dissemination together with the product release on the market.

ACKNOWLEDGMENTS

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, *41*(3), 1–52. https://doi.org/10.1145/1541880.1541883

Cichy, C., & Rass, S. (2019). An overview of data quality frameworks. *IEEE Access*, *7*, 24634–24648. https://doi.org/10.1109/ACCESS.2019.2899751

Feltham, G. A. (1968). The value of information. *The Accounting Review*, 684–696.

Guoling, L., & Qinyun, W. (2008). *Research on e-business model of distance education*. 400–403. https://doi.org/10.1109/CSSE.2008.145

Haug, A., Zachariassen, F., & Van Liempd, D. (2011). The costs of poor data quality. *Journal of Industrial Engineering and Management*, 168–193. http://dx.doi.org/10.3926/jiem.2011.v4n2.p168-193

Kawa, A., & Pierański, B. (2021). Green logistics in E-commerce. *Logforum*, *17*(2), 1. https://doi.org/10.17270/J.LOG.2021.588

Lone, S., Harboul, N., & Weltevreden, J. (2021). *2021 European E-commerce Report*.

Marsh, R. (2005). Drowning in dirty data? It's time to sink or swim: A four-stage methodology for total data quality management. *Journal of Database Marketing & Customer Strategy Management*, *12*, 105–112. https://doi.org/10.1057/palgrave.dbm.3240247

Niemir, M., & Mrugalska, B. (2021). Basic Product Data in E-Commerce: Specifications and Problems of Data Exchange. *EUROPEAN RESEARCH STUDIES JOURNAL*, *XXIV*(Special Issue 5), 317–329. https://doi.org/10.35808/ersj/2735

Niemir, M., & Mrugalska, B. (2022). *Product Data Quality in e-Commerce: Key Success Factors and Challenges*. 13th International Conference on Applied Human Factors and Ergonomics (AHFE 2022). https://doi.org/10.54941/ahfe1001626

Qalati, S. A., Vela, E. G., Li, W., Dakhan, S. A., Hong Thuy, T. T., & Merani, S. H. (2021). Effects of perceived service quality, website quality, and reputation on purchase intention: The mediating and moderating roles of trust and perceived risk in online shopping. *Cogent Business & Management*, *8*. https://doi.org/10.1080/23311975.2020.1869363

Reeves, C., & Bednar, D. (1994). Defining quality: Alternatives and implications. *Academy of Management Review*, 419–445. https://doi.org/10.2307/258934

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 86–95. https://doi.org/10.1145/240455.240479

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *J. Manage. Inf. Syst.*, *12*(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

Zhou, C. H., Chen, B., Gao, Y., Zhang, C., & Guo, Z. J. (2011). *A technique of filtering dirty data based on temporal-spatial correlation in wireless sensor network*. 511–516. https://doi.org/10.1016/j.proenv.2011.09.083

Zmud, R. (1978). An empirical investigation of the dimensionality of the concept of information. *Decision Sciences*, 187–195. https://doi.org/10.1111/j.1540-5915.1978.tb01378.x

Żuchowski, W. (2022). The smart warehouse trend: Actual level of technology availability. *Logforum*, *18*(2), 7. https://doi.org/10.17270/J.LOG.2022.702

Maciej Niemir ORCID ID: - https://orcid.org/0000-0002-1054-4285
Łukasiewicz - Poznań Institute of Technology,
Faculty of Engineering Management,
Poznan University of Technology, Poznan, **Poland**
e-mail: maciej.niemir@pit.lukasiewicz.gov.pl

Beata Mrugalska ORCID ID: - https://orcid.org/0000-0001-9827-9449
Faculty of Engineering Management,
Poznan University of Technology, Poznan, **Poland**
e-mail: beata.mrugalska@put.poznan.pl